Computing similarity between text and human summaries. A comparison of models.

Marianela Crissman
Northeastern Illinois University

Abstract. When tutoring, the ability to compare user responses with the correct answer is fundamental. With a view to contributing to research on Intelligent Tutoring Systems, this study has focused on the development of models that let us compute comparisons between two texts. Using scikit-learn library and python, we constructed a set of modules to generate eighteen (18) different vector space models (VSM); each of which could potentially be used to perform the comparison. Each model would differentiate itself from another based on: usage of stopwords; type of scoring when vectorizing the document (frequency, zero-one or TFIDF); and computation of Latent Semantic Analysis (LSA) using 100 and 300 dimensions during singular value decomposition (SVD). In order to evaluate each model, a testing dataset was created where 10 different paragraphs (LP1-LP10) were extracted from Mi Guia (Iacobelli et al 2018), an existing application for breast cancer education; and on each paragraph, well educated adult volunteers derived short sentences. Such sentences were either a summary or paraphrasing of the corresponding paragraph. Our test data set contained ten tasks, and each task was formed by a paragraph (LPx) and 6 sentences (SP(i,x), i=1..6) that needed to be ranked. We computed a machine ranking on each task using different models, and at the same time obtained human's perception using Amazon Mechanical Turk. We used different ranking correlation methods to determine how similar these two rankings were: the human and machine. Finally, we computed an average correlation per model and a standard deviation to determine which models yielded more human-like responses. Even though LSA with 300 dimensions, stop words and zero-one weights showed better performance, there are many other variables that need to be further studied in order to more confidently select a model.

Keywords: text similarity, vector space modeling, latent semantic analysis, singular value decomposition.

1 Introduction

For a very long time, schools and governments have been trying to increase the number of educated people. This certainly brings improvement to their communities. However in many countries, especially in the USA, education can be very expensive and many times unreachable for people that have potential.
In 2012 Coursera was founded by computer science professors, the goal behind it was to deliver educational content so everybody who wanted to learn was able to do it. Other online tools, such as Udacity, Udemy, EdX, Khan Academy and many more, emerged with the same goal. Youtube was also very popular hosting tutorials on many different subjects. These were great initiatives, however all of them were mostly a one-way communication type

of learning. Should the learner have questions, it would take a long time to get answers or feedback; quizzes were modified to be multiple selection because grading was fast and easy (automated). What about critical thinking? Questions that required answers like paragraphs?.
Intelligent Tutoring Systems, ITS, are complex systems that can help and guide people on their learning journey. However in order to better do this job they need to be able to identify when two texts are similar; in other words the systems has to compare natural language user-responses to natural language textbook responses. In this study, we discuss ways of comparing two texts.

2 Background

There are several ways to successfully compare two texts. Landauder (2013) provides several examples on how two texts can be compared; mostly using cosine similarity. For which it is needed to have two vectors to be compared.
Any collection of documents or corpus can be represented by a term-document matrix, where each term in the corpus corresponds to one row, and each column to one document of the corpus. Every entry cell is a functional representation of the word in the document. Such functions can be any of the following: term frequency, binary(zero-one), inverse document term-document frequency, entropy, among others. Each row or column can be treated as vectors.
Latent semantic indexing (LSI) or latent semantic analysis (LSA): a theory and method for extracting and representing the contextual-usage meaning of words which is achieved by statistical computations applied to a large corpus of text (Landauer et al 1998). LSA rests on a single constraint, the representation of a meaningful passage must be composed as a function of the representations of the words it contains (Landauer et al 2013). In other words, LSA assumes that a text is a representation through some function or processing of the words it contains.

$$passage\ or\ document\ =\ \sum(\ word_1,\ \ word_2,\ \ word_3\ \ldots\ \ word_n)$$

LSA uses a training corpus, which is a collection of passages or documents. Each document can be represented with the equation above. This example illustrates very well the idea behind LSA (example taken from Landauer 2013): if we have two equations $A\ +\ 2B\ =\ 8$ and $A\ +\ B\ =\ 5$, none of them (individually) tells us what the value of A and B is, however when we combine them it becomes possible for us to know what A and B are. Therefore, each document is a representation of the words it contains. Constructing a system of equations where each row corresponds to a unique word and every column to a document of the training corpus.
This brings to attention that the idea that adding or removing documents would, in theory, modify our system of equations, hence the meaning of a word. This is why, successful simulation of human passage or word meaning strongly depends on having sufficient large training corpus. This system of equations or matrix is said to hold word vectors: a row containing several elements. It is suggested that such vectors have between 200 and 500

elements (also called dimensions). For all we are concern one word might have around 300 close relations with other words, ie. cat, dog and cow, they are all animals; cow, dalmatian and zebra are all animals but also have black and white skin colors. We could say that one relation is related to a dimension.
Landauer & Dumais et al (1997) attempted to answer this question “how can people know as much as they do with as little information as they get”. They used LSA as a method to establish more solid relations between words. They showed a small example using a term-document matrix, generated with 9 different texts. Landauer & Dumais et al (1998) use count frequency for values in the term-document matrix.
Landauer & Dumais, explain in multiple publications how they apply a mathematical technique to decompose the term-document matrix into three factors, followed by a dimensionality reduction, and later a reconstruction of the new matrix. This is the so-called Singular Value Decomposition (SVD). This is an approximation of the original. We can use such approximations to extract latent information about the terms or documents, for example a relation between two words that was not obvious, after applying LSA can become more obvious.
They computed the cosine similarity between two words: *human-user* and *human-minor*. Their similarity increased from *-0.38* to *0.94* and *-0.29* to *-0.83*, respectively. That study performed a test, when carrying out SVD, to try to identify what number of dimensions yielded better results. There is no correct or proven number yet, however good results have been found using between 100 and 400. Deerwester, Landauer et al (1999) []
have conducted studies using 100 dimensions and have obtained decent results.

3 Method

For this study, eighteen vector space models (VSM) were generated. Such models can be categorized as: term-document vectors and LSA. For each model we variated several parameters which resulted in a variety of models. Our models contained around 70,000 different words that are widely used within an English speaking population.

With each one of these models, we created sentence vectors: a vector representation of the words contained in the sentence which will be utilized for comparing texts. In order to achieve this, the two sentences to be compared would have to go through this vectorization process. Once we have the vector representation of each sentence, the system will compute the similarity between them and output the result.

In order to evaluate how well the model performed, ten different paragraphs were extracted from *Mi Guia* (Iacobelli et al 2018), a guide that contains information about Breast Cancer; for each of the ten paragraphs, six educated volunteers generated a sentence for each paragraph. These are individuals who have completed higher education, five of them have ages between 29 and 33 years old, with the exception of one who is in her 60s; all of them either live in an English speaking population or have spoken the language long enough to

comprehend texts. Each of them was required to write what they understood about each paragraph, sometimes using paraphrasing, and others, all or some of the main ideas.

The goal of this study was for US-based humans to rank the sentences based on their similarity compared to the paragraph. This was achieved via Amazon Mechanical Turk. Finally, results from Amazon Mechanical Turk (ranking by humans) and from our system (ranking by machine) were obtained and correlated using Spearman's ranking algorithm.

Several modules were created in order to carry out the different tasks needed to evaluate the similarity between texts:

- Generating the model
- Using the model
- Evaluating the model

3.1 Generating the Vector Space Models

Data

A corpus of senate speeches (Diermeier et al. 2012) was chosen because in a senate speech, there is a large diversity of vocabulary used, involving a wide variety of topics. We consider this to be a good enough representation of lexicon and topics that are popular in an English speaking population's speech. This corpus contained approximately 136,000 different documents and around 115,000 different words; its size was 758MB and had an XML-like format.

Before the creating the vector space model, the collection of documents was prepared according to the following steps:

- Removing punctuation
- Converting the text to be lowercase
- Replacing four-digit numbers with '@*YEAR*@'
- Replacing digit numbers prefixed by currency symbols: $, £, € with '@*PRICE*@'
- Ignoring words with a low document frequency by keeping words that appeared in at least 2 of the documents.

This preprocessing was intended to have a better set of words, ignoring mostly digits and very obscure words, which resulted in a reduction of 40,000 terms on our matrix model. Such terms were street numbers, distance measures, ordinal numbers, typos and very rare words.

Vector Modeling

Scikit-learn library and Python were used to generate the term by document matrix, where each column represented a document of the corpus and each row represented a unique word: unigram. Our system defined a word as a group of at least two letters. Every cell in

the matrix contained a score for a given word in a given document. We created three term-document matrices. Every matrix was scored according to one of the following techniques:

- Count: represented the number of times the word appeared in the document identified by the column.
- Zero-One or Binary: represented with a one (1) or a zero (0) the existence of the word in the document identified by the column.
- TF-IDF: a more complex number that represented the importance of the word in the document identified by the column. In other words, the weight assigned to each token/cell not only depends on its frequency in a document but also how recurrent that term is in the entire corpora. The lower the number the less influence the word had on the document.

When creating the scripts, in order to save memory space on the machine, every cell was considered to be of data type: 32-bit float.

Another parameter that we modified to obtain a new type of model was the use of stop words. This parameter increased our model count to six (6). Stop words are words that are commonly filtered out in texts when processing natural language data; words like "*the"*, *"a"*, *"of"* etc, which appear very frequently in texts but that do not add significant value to the text. So, this process removes any non-contributing word that is found on the corpus. This is a step that takes place before generating the term-document matrix.

Latent Semantic Analysis

An alternative to the VSM previously mentioned would be using LSA, for this a term-document matrix was required. In order to apply LSA to our existing VSMs, we considered the dimensionality reduction when performing SVD. LSA modifies the term-document matrix such that words that co-occur together would have similar vectors. Even if those words did not appear in the same texts. The steps we used to perform LSA were:

- Decomposed the term-document matrix into its factor matrices (Landauer et al 1998).
- Reduced dimensionality of factor matrices: 100 or 300.
- Multiplied the dimensionality reduced matrices to reconstruct an approximation of the original matrix.

We have described the **eighteen** vector space models that were used to compare texts:

Table 3.1.1: Our vector space models and their variations.

VSM	dimensionality	score	stop words
1	-	count	yes
2	-	count	-

3	-	tfidf	yes
4	-	tfidf	-
5	-	zero-one	yes
6	-	zero-one	-
7	100	count	yes
8	100	count	-
9	100	tfidf	yes
10	100	tfidf	-
11	100	zero-one	yes
12	100	zero-one	-
13	300	count	yes
14	300	count	-
15	300	tfidf	yes
16	300	tfidf	-
17	300	zero-one	yes
18	300	zero-one	-

In order to optimize and at the same time avoid memory issues on the machine, all VSMs that used LSA were truncated, keeping all the words but only the first 1000 dimensions. For those VSMs that did not use LSA, all words and dimensions (columns) were kept and a sparse matrix was stored instead. This format considerably reduced the memory space required per model: from ~37GB to ~300MB; and the computation space and time, per model: from ~6 minutes with ~62GB to under 1 minute with a few GBs.

3.2 Testing Dataset

We created our test dataset by including ten paragraphs and six sentences attached to each paragraph. Each paragraph, which we identify by LP1-LP10, was extracted from *Mi Guia []*. The six sentences related to each paragraph were generated by well educated volunteers.

Our volunteers belong to an adult population of ages 29 to 33 years, with the exception of one who is in her 60s; each of them have completed higher education studies in disciplines such as engineering, medicine and biology; half of the volunteers have spoken English all their lives, the other half has spoken it long enough to maintain a conversation and write reports.

We created a form containing all ten paragraphs (LP1-LP10). We distributed this form to the volunteers. They were required to read each one of the paragraphs and write a short text either summarizing or paraphrasing the paragraph. Answers with all or some of the main ideas were acceptable. We suggested these answers were written in one or two sentences.

For each paragraph (ie. LP1) we took four sentences that were derived from it, and two other sentences that were derived from different paragraphs (ie: LP10 and LP9). The decision was made so we had some control over the human rankings, and a way to more easily test rankings from our system. Every group of one paragraph and 6 sentences we called a task to be ranked. Like the following table shows.

Table 3.2.1. Example of a task to be ranked.

Paragraph: LP1	Cancer begins in cells. Cells are the building blocks that make up all parts of our bodies, including our breasts. Cancer begins when the cells in the breast change and grow into a tumor. If not removed or treated, a breast cancer tumor can spread to other parts of the body and become lethal. These days, there are many treatment options for breast cancer and the majority of women are cured of breast cancer.
sentence: SP(1,1)	Systemic therapy reduces the chances of breast cancer coming back.
sentence: SP(2,1)	Mastectomy is a procedure that eliminates the breast.
sentence: SP(3,1)	Cancer starts in the smallest entity of our bodies; the cell grows too the parts of the body and if left untreated will probably end in death.
sentence: SP(4,1)	Cancer is a pathologic function of cells, your body is made by cells. This means that you can have cancer in a specific part of your body or you can have cancer everywhere. When it starts, it is usually in a determined territory and it is easier to treat and get a better prognosis in that moment.
sentence: SP(5,1)	Cancer starts in the body cells and when it grows out of control, it becomes a tumor. For breast cancer there are actually several treatment options that can prevent the spreading of malignant cells to other body organs, so a big number of women are cured from this disease.
sentence: SP(6,1)	Sometimes cells in the body can change, and may grow into cancerous tumors, which must be removed from the body before they become lethal.

This table shows an example where paragraph LP1 explains what is cancer and how it starts on the body; sentences identified by SP(3,1), SP(4,1), SP(5,1) and SP(6,1) talk about that, however it is obvious that sentences SP(1,1) and SP(2,1), even though they talk about breast cancer, clearly were not derived from the LP1 paragraph.

Our test data set had ten tasks to be ranked (like the one above), and ideally we would expect the non-related sentences to be ranked at the very end.

3.3 Applying the Model

We developed a module to apply rankings on the test dataset using each VSM. This module produced a machine ranking on each task of the test dataset. In order to produce one ranking, there were six comparisons that needed to be done. Each comparison would compute the cosine similarity between the paragraph (ie. LP1) and the sentence (ie. SP1). So, there were six comparisons per task, ten tasks per dataset, and eighteen models applying rankings to the dataset. This yielded a total of 1080 comparisons.

Comparison process

The way a comparison was performed is as follows, having two test sentences *LPx* and *SP(i,x)*:

1. Clean the sentences: *LPx*, *SP(i,x)*
    - Removing punctuation,
    - Transforming the text to lowercase,
    - Replacing all 4-digit numbers to @*YEAR*@ and all $-4-digit to @*PRICE*@.
2. Vectorize sentences: converting the sentence to a vector
    - Obtaining the vector corresponding to each word in the sentence, even if there were dulipcates. In case a word did not exist on the VSM, a default vector filled with 0.00001 was returned.
    - Performing an average on all vectors from the previous step.
3. Compute cosine similarity with the two sentence vectors.

After all six sentences were compared to the corresponding paragraph, we sorted the results by their cosine similarity result, the far away from 1 the less similar were. These results were ranked with numbers from 1 to 6: 1 meant the sentence is most similar and 6 meant less similar.

3.4 Using Amazon Mechanical Turk

In order to evaluate how well the system computed similarities we needed to have a reference or comparison that was already “correct”. This is what we aimed to achieve by using Amazon Mechanical Turk.

From the dataset previously shown, ten files were constructed to launch surveys using Amazon Mechanical Turk (AMT). Each file contained one paragraph (*LPx*) and six sentences (*SPx*) to be ordered, which we call: a human intelligent task (HIT). We launched ten assignments per HIT, that is, for each task we expected to have ten AMT worker responses. All AMT workers were given 4 minutes to order the sentences and submit; upon completion and approval on our side, the reward was issued. All workers were required to be located in the United States.

In addition, a simple web interface was developed in order to let the AMT worker easily and intuitively order the sentences. Submitted results were available for download on the AMT portal.

3.5 Evaluating the Model

Results from AMT portal were downloaded. One HIT had responses from ten workers, that is, ten possible human rankings. On each HIT response, we computed the average ranking, where each of the six sentence positions were recalculated using the average on the orders given by the workers. Following the procedure explained above, we established the final human ranking per task (LP1-LP10).
On the other hand, our machine rankings were already computed and ready to be used. We compared the human rankings with each of the eighteen VSM rankings by calculating: Spearman's rank-order correlation coefficient, Kendall Tau correlation and *Order Preservation Measure*. This last algorithm was designed for the purpose of this study.

Spearman rank-order correlation coefficient is a good fit for datasets whose variables contain ordinal values []. An ordinal variable is similar to a categorical variable; the difference between the two is that there is a clear ordering of the variables []. Kendall Tau correlation [] is another approach to comparing two ranked datasets. Typically used when the dataset cannot comply with the two assumptions that Spearman's require.

Order Preservation Measure is an algorithm that compares how well the order between two categorical orderings rank A and rank B is preserved, being B the correct order. Let's have $A = [\, Illinois,\ Chicago,\ Edgewater,\ USA\,]$ and $B = [\, USA,\ Illinois,\ Chicago,\ Edgewater]$, where the first position has a degree of importance higher than the last position.

Table 3.5.1. Tagging process for PreserveOrder algorithm.

B	1-USA	2-Illinois	3-Chicago	4-Edgewater
A	Illinois	Chicago	Edgewater	USA
A_numerical	2	3	4	1

Knowing this, we proceeded to assign each element of the A_numerical list with the corresponding index (starting at 0 or 1) based on B's rank, Table 3.5.1 shows the A_numerical list. Then we proceed to check if the elements of A_numeical are in ascending order; that is, letting *i* and *j* be different indices of A_numerical and *i* < *j*, the element in index *i* should be smaller than the element in index *j*. There are a total of *N* (n Choose 2) *isAscending* comparisons, *K* of which result in 1; then our resulting value is the quotient between *K* and *N*.

These equations illustrate the mathematical computations:

$$orderPreservationMeasure = \frac{1}{C_{n,2}} \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} isAscending(i,j,A_{num}),$$

$$isAscending(i,j,A_{num}) = 1 \quad if\ A_{num}[i] < A_{num}[j], \quad 0\ otherwise.$$

4 Results

The specification of the machine used to compute results in this study are:

Processor: 2.7 GHz Quad-Core Intel Core i7
Memory: 16 GB 2133 MHz LPDDR3
Graphics: Intel Iris Plus Graphics 655 1536 MB

4.1 Generating a Model Results

- Processing the corpus before vectorization helped reduce the number of terms by 39.2%.
- Loading the corpus, a text file of size 758MB, took between 17 to 20 seconds.
- Term by document matrix took between 40 seconds and 60 seconds to be generated.
- Dimensionality reduction took between 30 seconds and 210 seconds to be computed, such measure depended on the number of dimensions selected.
- Storing non-LSA vector space models on disk took less than a second, however for LSA models it took approximately 2 minutes, this number depends on the size of the model to be stored.

The system was able to generate one vector space model in under 5 minutes. The resulting model was stored in a file that did not exceed ~300 MB, however its use required approximately 35GB of memory to process similarity results.

4.2 Similarity Results

The following table shows cosine similarity between two pairs of sentences that were mentioned on Landauer et al. (2013) work. We took the same sentences and produced corresponding results per model.

Pair A: should be somewhat similar - 0.66 (Landauer et al 1998)

- *Several doctors operated on a patient.*
- *The surgery was done by many physicians.*

Pair B: should not be similar - 0.03 (Landauer et al 1998)

- *A circle's diameter.*

- *Music of the spheres.*

Pair C:

- *In America, you can always find a party.*
- *Members of America joined the Russian Party.*

Table 4.2.1. Cosine similarity on Pairs: A and B.

Vector Space Model	Cosine Similarity		
	Pair A	Pair B	Pair C
lsa:- dim:- sco:count sw:yes	0.359	0.048	0.833
lsa:- dim:- sco:count sw:-	0.945	0.945	0.950
lsa:- dim:- sco:tfidf sw:yes	0.220	0.034	0.724
lsa:- dim:- sco:tfidf sw:-	0.860	0.870	0.882
lsa:- dim:- sco:zeroone sw:yes	0.336	0.079	0.814
lsa:- dim:- sco:zeroone sw:-	0.911	0.925	0.931
lsa:yes dim:100 sco:count sw:yes	0.918	0.379	0.878
lsa:yes dim:100 sco:count sw:-	0.939	0.941	0.946
lsa:yes dim:100 sco:tfidf sw:yes	0.942	0.449	0.838
lsa:yes dim:100 sco:tfidf sw:-	0.865	0.887	0.901
lsa:yes dim:100 sco:zeroone sw:yes	0.950	0.410	0.797
lsa:yes dim:100 sco:zeroone sw:-	0.965	0.953	0.965
lsa:yes dim:300 sco:count sw:yes	0.874	0.354	0.812
lsa:yes dim:300 sco:count sw:-	0.939	0.941	0.945
lsa:yes dim:300 sco:tfidf sw:yes	0.901	0.311	0.715
lsa:yes dim:300 sco:tfidf sw:-	0.857	0.877	0.898
lsa:yes dim:300 sco:zeroone sw:yes	0.929	0.387	0.786
lsa:yes dim:300 sco:zeroone sw:-	0.929	0.943	0.952

Green highlights are expected values, red highlights are non expected values. Each vector space model can be read the following way:

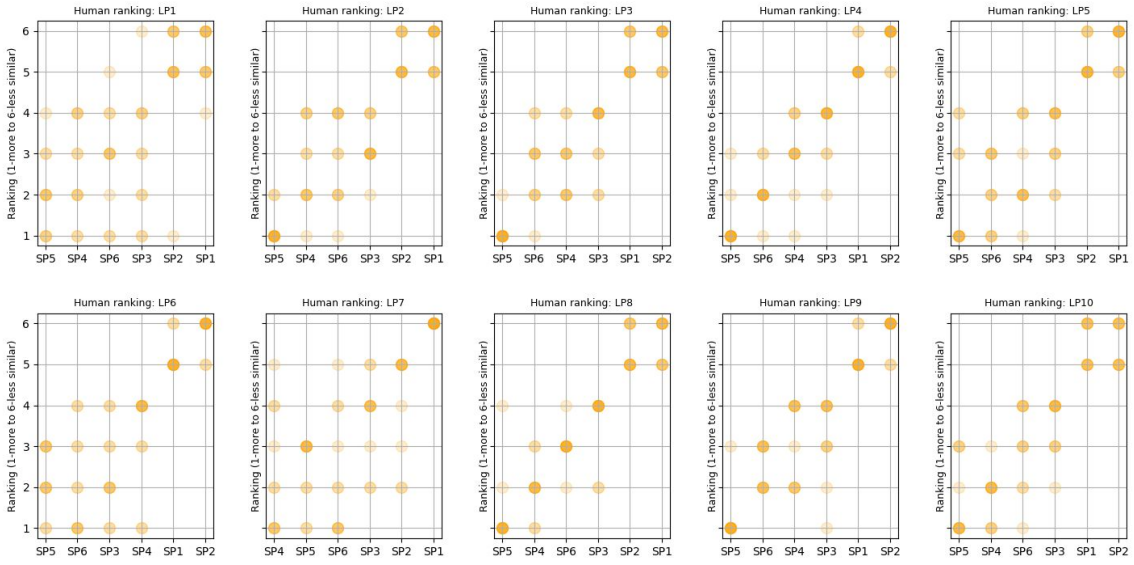
- lsa: whether LSA technique was used or not
- dim: number of dimensions for SVD process, 100 or 300
- sco: the function used to fill the cells on the term-document matrix count, tfidf zero-one
- sw: whether stop words were removed during corpus preprocessing.

It took approximately 90 minutes to get machine rankings on each of the tasks using all 18 models.

4.3 Human Results

The following graphs show the workers' ranking on each task (LP1-LP10).

Figure 4.3.1. AMT Workers ranking.



Each graph shows the sentences (SP1-SP6) ordered from most similar (left) to least similar(right), and we used them to label the horizontal axis. On the vertical axis we show the ranking selected by a human. Different colors on the graphs represent the different workers who completed the rankings.

4.2 Evaluation the Model Results

The following table shows the Spearman's Correlation. Each row represented the vector space model used to compute the similarities, and each column represented every ranked task from the test data set. Cells highlighted in different shades of blue suggest the darker the blue the higher the correlation; conversely, the lighter the shade the lower the correlation. Results like the table and chart below were also generated using Kendall Tau Correlation and our Order Preservation Measure algorithm. They can be found on Appendix C.

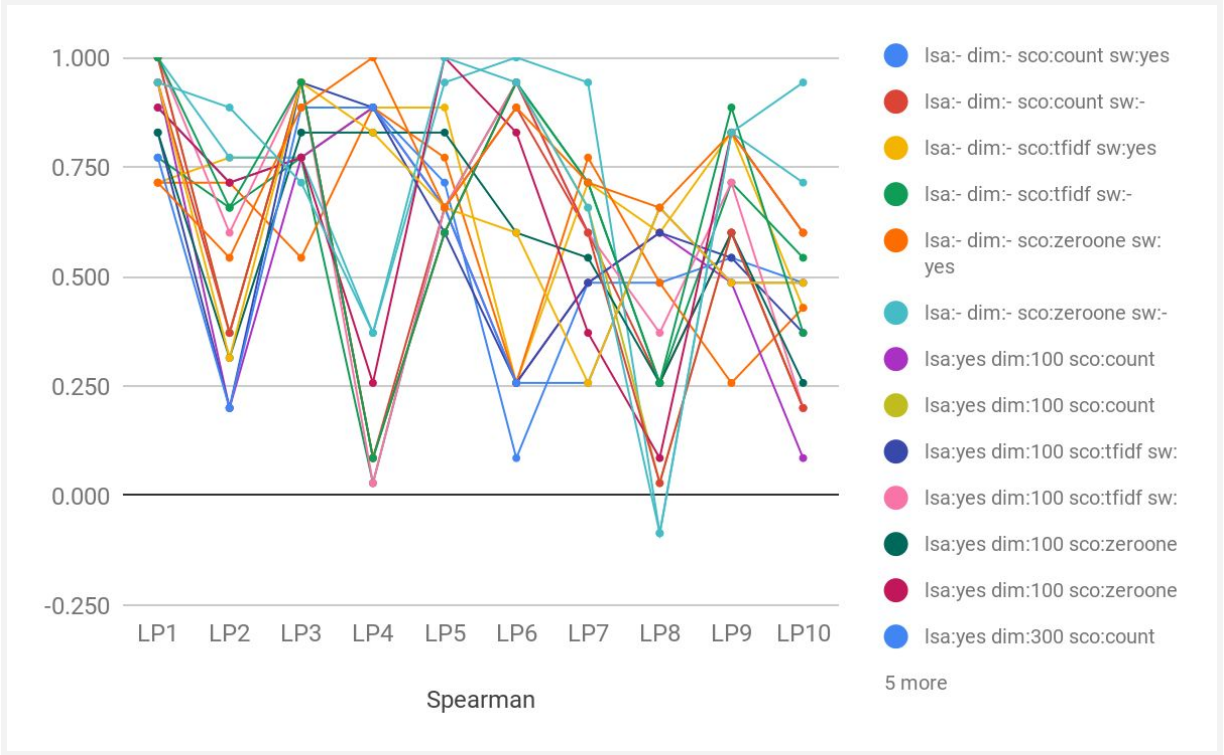
Table 4.2.1: Spearman's Correlation per model per task.

Spearman	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8	LP9	LP10
lsa:- dim:- sco:count sw:yes	0.886	0.714	0.771	0.886	0.714	0.086	0.486	0.486	0.543	0.486
lsa:- dim:- sco:count sw:-	0.943	0.371	0.943	0.086	0.657	0.886	0.600	0.257	0.600	0.200
lsa:- dim:- sco:tfidf sw:yes	0.714	0.771	0.771	0.886	0.886	0.257	0.714	0.600	0.829	0.429
lsa:- dim:- sco:tfidf sw:-	0.771	0.657	0.771	0.029	0.657	0.943	0.714	0.257	0.714	0.543

lsa:- dim:- sco:zeroone sw:yes	0.714	0.714	0.543	0.886	0.771	0.257	0.771	0.486	0.257	0.429
lsa:- dim:- sco:zeroone sw:-	1.000	0.771	0.771	0.371	0.943	1.000	0.943	-0.086	0.829	0.714
lsa:yes dim:100 sco:count sw:yes	0.943	0.200	0.771	0.886	0.657	0.257	0.486	0.600	0.486	0.086
lsa:yes dim:100 sco:count sw:-	1.000	0.371	0.943	0.086	0.600	0.943	0.657	0.029	0.600	0.200
lsa:yes dim:100 sco:tfidf sw:yes	0.829	0.200	0.943	0.886	0.600	0.257	0.486	0.600	0.543	0.371
lsa:yes dim:100 sco:tfidf sw:-	1.000	0.600	0.943	0.029	0.657	0.943	0.600	0.371	0.714	0.200
lsa:yes dim:100 sco:zeroone sw:yes	0.829	0.314	0.829	0.829	0.829	0.600	0.543	0.257	0.600	0.257
lsa:yes dim:100 sco:zeroone sw:-	0.886	0.714	0.771	0.257	1.000	0.829	0.371	0.086	0.829	0.600
lsa:yes dim:300 sco:count sw:yes	0.771	0.200	0.886	0.886	0.657	0.257	0.257	0.657	0.486	0.486
lsa:yes dim:300 sco:count sw:-	1.000	0.371	0.943	0.086	0.600	0.943	0.600	0.029	0.600	0.200
lsa:yes dim:300 sco:tfidf sw:yes	0.943	0.314	0.943	0.829	0.657	0.600	0.257	0.657	0.486	0.486
lsa:yes dim:300 sco:tfidf sw:-	1.000	0.657	0.943	0.086	0.600	0.943	0.714	0.257	0.886	0.371
lsa:yes dim:300 sco:zeroone sw:yes	0.714	0.543	0.886	1.000	0.657	0.886	0.714	0.657	0.829	0.600
lsa:yes dim:300 sco:zeroone sw:-	0.943	0.886	0.714	0.371	1.000	0.943	0.657	-0.086	0.829	0.943

Another alternative to visualizing correlation results is shown on the following figure, each line corresponding to one model. The vertical axis displays the correlation number, and the horizontal axis shows for which tasks we obtain such correlation values.

Figure 4.2.1 Chart: Spearman's Correlation per model per task.



In order to determine which module yielded better results, we computed an average on the correlations with their standard deviation. The higher the average and the smaller the standard deviation, the better the model.

Table 4.2.2. Average and Standard Deviation per model

	Spearman		Kendall		Order P Measure	
	Average	**Standard Deviation**	**Average**	**Standard Deviation**	**Average**	**Standard Deviation**
lsa:- dim:- sco:count sw:yes	0.606	0.241612	0.467	0.208463	0.733	0.104231
lsa:- dim:- sco:count sw:-	0.554	0.315064	0.480	0.284236	0.740	0.142118
lsa:- dim:- sco:tfidf sw:yes	0.686	0.203818	0.533	0.201231	0.767	0.100615
lsa:- dim:- sco:tfidf sw:-	0.606	0.269979	0.507	0.259534	0.753	0.129767
lsa:- dim:- sco:zeroone sw:yes	0.583	0.222213	0.427	0.189085	0.713	0.094542
lsa:- dim:- sco:zeroone sw:-	0.726	0.340521	0.627	0.359973	0.813	0.179986
lsa:yes dim:100 sco:count sw:yes	0.537	0.290687	0.413	0.296148	0.707	0.148074
lsa:yes dim:100 sco:count sw:-	0.543	0.360398	0.480	0.358047	0.740	0.179023
lsa:yes dim:100 sco:tfidf sw:yes	0.571	0.255905	0.427	0.267037	0.713	0.133518
lsa:yes dim:100 sco:tfidf sw:-	0.606	0.324872	0.533	0.328107	0.767	0.164054
lsa:yes dim:100 sco:zeroone sw:yes	0.589	0.242137	0.480	0.221777	0.740	0.110889
lsa:yes dim:100 sco:zeroone sw:-	0.634	0.300204	0.520	0.275412	0.760	0.137706
lsa:yes dim:300 sco:count sw:yes	0.554	0.258093	0.440	0.241804	0.720	0.120902
lsa:yes dim:300 sco:count sw:-	0.537	0.358834	0.467	0.355556	0.733	0.177778
lsa:yes dim:300 sco:tfidf sw:yes	0.617	0.239501	0.507	0.243838	0.753	0.121919
lsa:yes dim:300 sco:tfidf sw:-	0.646	0.317359	0.573	0.306634	0.787	0.153317
lsa:yes dim:300 sco:zeroone sw:yes	0.749	0.145562	0.613	0.182709	0.807	0.091355
lsa:yes dim:300 sco:zeroone sw:-	0.720	0.340148	0.613	0.352487	0.807	0.176243

Average correlation results with high and low values were highlighted in green and red, respectively. Small values of standard deviation were highlighted in yellow. We do this to better visualize, on average, which model provided results that were very close to the human ranking. We understand that by taking the average of all correlations (belonging to one model) we could be choosing a model with very dispersive correlation values, this is why we look at the standard deviation, so we can see how consistent the model was.

The following graph shows the human average and machine ranking, only on one of the models (all graphs can be found in Appendix B). The vertical axis shows the rank given. The

horizontal axis shows sentences ordered by the humans, from more similar to less similar. Each subplot corresponds to the responses per task.

Figure 4.2.2. Human-Machine rankings. Best results.

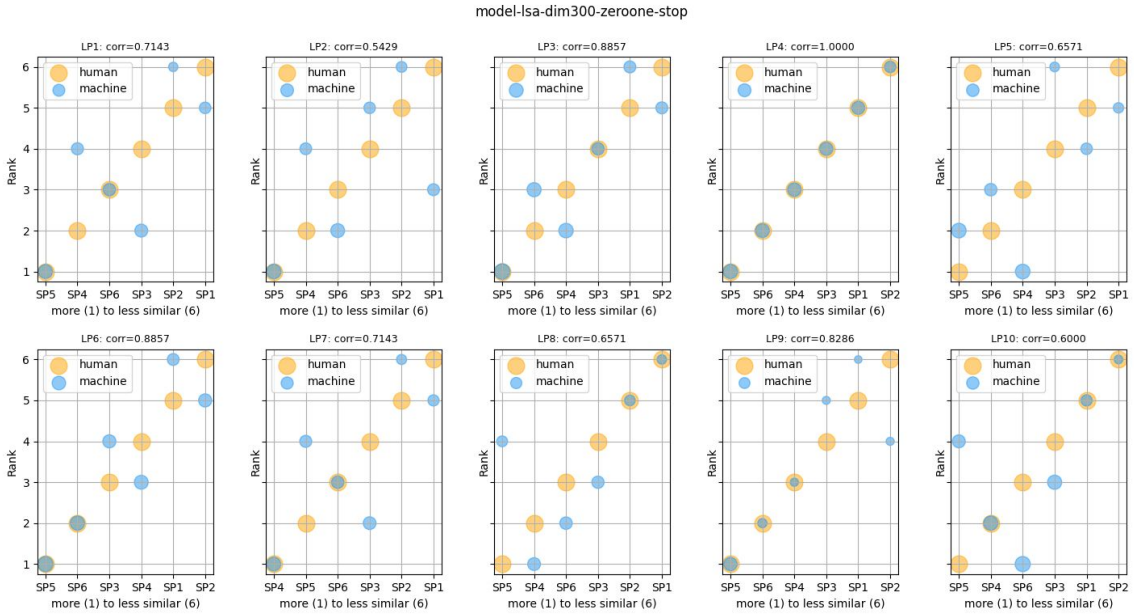
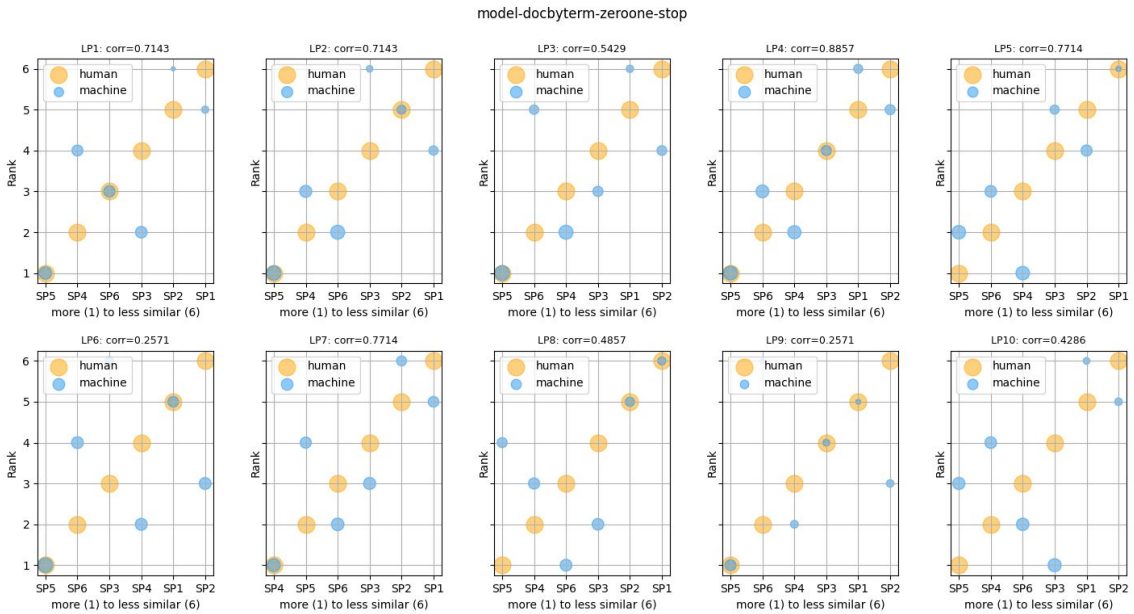


Figure 4.2.3. Human-Machine rankings. Worst results.



5 Discussion

We can observe from Figure 4.3.1 that even for humans it was difficult to agree on the ranking of a specific task. It was clear that sentences S1 and S2 were not so hard to rank with values 5 or 6. They were less similar to the human eye. Such results were expected. As for the rest of the sentences, there was a lot of discrepancy on their ranking. This might be due to the fact that the sentences S3-S6 were very similar among themselves, therefore it was more challenging to rank them.

We used three measures to determine how well the rankings were computed by the machine. From table 4.2.1 and figure 4.2.1, we can observe that results were very satisfactory on tasks LP1 and LP3, however it was shocking to see such drastic change for tasks LP8 and LP10 where none-to-little relation was found between the machine ranking and human ranking. Finally, we observed tasks LP4 and LP6. It is interesting how the correlation alternated between 0.02 and 0.94 depending on the model's usage of stop words.

Table 2.2.2, shows the average and standard deviation on each of the measures used. The model that consistently yielded high correlations was ***lsa:yes dim:300 sco:zeroone sw:yes***. Conversely, the model that consistently yielded low correlation values was ***lsa:- dim:- sco:zeroone sw:yes***.

6 Conclusion

When forming a definite human ranking there were a lot of discrepancies in the human responses. This might have been caused by the fact that the sentences (S3-S6) were very similar among themselves; or to the level of comprehension of the population who was ranking the tasks. It would be recommended for a larger population to rank each task. Another improvement would be to have a Linguistic team revise the test data set, ensuring that the sentences have more noticable differences. This might improve the results of this study.

During similarity computation on task LP8 (talked about Lumpectomy), words like *nipple, areola, pathologic, comorbidities* and *conservate* were not found in any of the models. Such words might have had a solid impact on the similarity comparison results. This task in particular, had short sentences that used a very similar vocabulary. Knowing that the vectorization in this study does not consider the order in which the words appear in a sentence, our models are going to consider these two sentences *"In America, you can always find a party"* and *"Members of America joined the Russian Party"* to be very similar. For future work it would be recommended using a corpus that included more medical terminology; and during vectorization, instead of unigrams, using n-gams.

We have seen that for a specific set of sentences, models can yield very opposite results depending on whether such models have used stop words. Even though we are not certain as to why such responses, it is an interesting fact to look at in further studies.

LSA can be very expensive on computation power, when dealing with a very large number of words and documents. This is one of its biggest challenges. The preprocessing of the corpus helped to reduce the number of rows, which helped reduce the size of the model. By how much more the model could be reduced if stemming and other natural language preprocessing techniques were applied is the question that remains unanswered.

We have seen that vector space models that use stop words and the binary (zero-one) function for every cell on the term-document matrix can improve if latent semantic analysis is applied.

7 References

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. Discourse Processes, 25, 259–284.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013) Handbook of latent semantic analysis.

Steinberger, J., & Ježek, K. (2012). Evaluation Measures for Text Summarization. COMPUTING AND INFORMATICS, 28(2). Retrieved from http://www.cai.sk/ojs/index.php/cai/article/view/37

Landauer, T. K. (2013). LSA as Theory meaning. Retrieved from http://cognaction.org/cogs105/readings/LSA.pdf

Iacobelli, Francisco and Adler, Rachel F. and Buitrago, Diana and Buscemi, Joanna and Corden, Marya E. and Perez-Tamayo, Alejandra and Penedo, Frank J. and Rodriguez, Melinda and Yanez, Betina R. (2018). Designing an mHealth application to bridge health disparities in Latina breast cancer survivors: a community-supported design approach. Design for Health vol.2 pp. 1--19

DANIEL DIERMEIER, JEAN-FRANÇOIS GODBOUT, BEI YU and STEFAN KAUFMANN
British Journal of Political Science
Vol. 42, No. 1 (JANUARY 2012), pp. 31-55

Spearman's Rank Order Correlation
https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php

LSA on Python - https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python

Cosine Similarity in Sci-Kit Learn - Retrieved from
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html

Vectorization in Sci-Kit Learn - Retrieved from
https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Amazon Mechanical Turk - Retrieved from https://www.mturk.com/

APPENDIX A

This is the dataset used to test the system.


```
[
 {
   "answer": "Cancer begins in cells. Cells are the building blocks that make up all parts of our bodies, including our breasts. Cancer begins when the cells in the breast change and grow into a tumor. If not removed or treated, a breast cancer tumor can spread to other parts of the body and become lethal. These days, there are many treatment options for breast cancer and the majority of women are cured of breast cancer.",
   "lp_id": "LP1",
   "possible_answers": [
     {
       "id": "SP1",
       "text": "Systemic therapy reduces the chances of breast cancer coming back. "
     },
     {
       "id": "SP2",
       "text": "Mastectomy is a procedure that eliminates the breast."
     },
     {
       "id": "SP3",
       "text": "Cancer starts in the smallest entity of our bodies; the cell grows too the parts of the body and if left untreated will probably end in death."
     },
     {
       "id": "SP4",
       "text": "Cancer is a pathologic function of cells, your body is made by cells. This means that you can have cancer in a specific part of your body or you can have cancer everywhere. When it starts, it is usually in a determined territory and it is easier to treat and get a better prognosis in that moment."
     },
     {
       "id": "SP5",
       "text": "Cancer starts in the body cells and when it grows out of control, it becomes a tumor.  For breast cancer there are actually several treatment options that can prevent the spreading of malignant cells to other body organs, so a big number of women are cured from this disease."
     },
     {
       "id": "SP6",
       "text": "Sometimes cells in the body can change, and may grow into cancerous tumors, which must be removed from the body before they become lethal."
     }
   ]
 },
 {
   "answer": "For the majority of women, breast cancer is not a death sentence. In the U.S., fewer women have been dying from breast cancer since the 1990s. Women whose breast cancer has not spread to other organs in their body like the lungs, liver, or brain, have over a 90% chance of being alive in 5 years. How long each person diagnosed with breast cancer will live and whether she will die from the disease depends on many things, including how much the cancer has spread throughout the body, the woman's overall health, and more.",
```

```
  "lp_id": "LP2",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "Cancer are malignant cells that grow into more cells to form a tumor which needs to be treated to not become lethal. "
    },
    {
      "id": "SP2",
      "text": "Systemic therapy is for the whole body and can target and kill cancer in other parts of the body."
    },
    {
      "id": "SP3",
      "text": "Depending on the advancement of breast cancer as well of other factors, the patient might survive."
    },
    {
      "id": "SP4",
      "text": "The perception of breast cancer is changing everyday. Women are not afraid because they know that if the diagnosis is early, they will have the opportunity to cure it depending on the comorbidities."
    },
    {
      "id": "SP5",
      "text": "For most of the women population who have been diagnosed with breast cancer, it is no longer a life threatening situation. Those whom their cancer has not spread to other organs like lungs, liver or brain have a 90% possibility of living the next five years or so."
    },
    {
      "id": "SP6",
      "text": "Breast cancer mortality depends on many factors, like health, lifestyle, and spread to other organs, but cancer is less deadly than 20-30 years ago."
    }
  ]
 },
 {
  "answer": "Research has not found that your attitude or level of stress causes cancer or makes cancer come back. It is normal to feel sad, angry, or frustrated sometimes and positive and hopeful and other times. People with a positive attitude may be more likely to stay social, which is important for cancer survivors. Getting physical activity through walking or exercise and emotional support are also helpful for women who had breast cancer.",
  "lp_id": "LP3",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "Breast cancer is not directly a death sentence. "
    },
    {
      "id": "SP2",
      "text": "Cancer in the breast, if not treated, can move to other parts of the body and cause death."
    },
    {
```

```
      "id": "SP3",
      "text": "Academically, there is no link between the stress level of a patient and the development of cancer."
    },
    {
      "id": "SP4",
      "text": "We cannot assure that if you have a positive attitude you can beat the cancer, but based on survivors experiences, it's important to keep close relationships so you can feel supported, and keep working out so you can feel healthy."
    },
    {
      "id": "SP5",
      "text": "There is no scientific evidence about the relation between stress, attitude and cancer. It is normal that a person can have different moods during their days, so sometimes they can feel sad, angry or frustrated; and other days happy and optimistic.  People with positive attitudes tend to be social which helps to avoid depression; also having someone to talk to about anything is very important for those who have overcome the disease."
    },
    {
      "id": "SP6",
      "text": "A stressful lifestyle or negative attitude cannot cause cancer, but being positive is correlated with helpful survival skills, such as socializing and exercising."
    }
  ]
 },
 {
  "answer": "So far herbal products have not been shown to cure cancer. However, there are some herbs that may help patients deal with the side effects of cancer treatments. We recommend that you talk to your doctor before taking any vitamins and herbal products as they might have an effect on your cancer treatments.",
  "lp_id": "LP4",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "Even though your attitude might not be a direct cause of cancer, it is helpful to stay positive and social. "
    },
    {
      "id": "SP2",
      "text": "There are many variables to determine whether a woman will die from breast cancer within the next five years."
    },
    {
      "id": "SP3",
      "text": "Herbal products and vitamins are not linked to any treatment for cancer, however, it can minimize the secondary effects of proper treatment."
    },
    {
      "id": "SP4",
      "text": "People are accustomed to consuming herbal products but there is no evidence of cure when they use them. However they could help with side effects of traditional cancer treatments."
    },
    {
```

```
      "id": "SP5",
      "text": "So far, there is no proof that herbal products heal cancer, but there are some that can help with the treatment side effects. Nevertheless any herbal or vitamin products that you might want to try should be advised by your treating physician."
    },
    {
      "id": "SP6",
      "text": "Herbal products may help with some cancer side effects, but you should discuss them with your doctor before starting to take any vitamins or supplements."
    }
  ]
 },
 {
  "answer": "Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control.\nIt's important to understand that most breast lumps are benign and not cancer (malignant). Non-cancerous breast tumors are abnormal growths, but they do not spread outside of the breast. They are not life threatening, but some types of benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a healthcare professional to determine if it is benign or malignant (cancer) and if it might affect your future cancer risk.",
  "lp_id": "LP5",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "Herbs might have beneficial effects on cancer treatments. "
    },
    {
      "id": "SP2",
      "text": "Getting good exercise, having a positive attitude, and letting your emotions out are good for those who have had cancer."
    },
    {
      "id": "SP3",
      "text": "The growth of malignant cells on the breast is called breast cancer, however, not all are malignant but can indicate a predisposition."
    },
    {
      "id": "SP4",
      "text": "Many types of breast tumor are benign. It's important to do a regular and frequent medical check up to figure out if it's malignant."
    },
    {
      "id": "SP5",
      "text": "Cancer begins when body cells begin to grow out of control. This can happen in any part of the body, but when this happens in the breast tissue this is called breast cancer.  Any breast lump or masses should be checked by a specialist to determine if it's malignant."
    },
    {
      "id": "SP6",
      "text": "A lump or change in the breast may be cancer or it may not be -- those non cancerous lumps are not damaging to your health, but only a doctor can tell what kind of lump it is."
    }
  ]
```

```
},
{
  "answer": "For most types of cancer, doctors use staging information to help plan treatment and to predict a person's outlook (prognosis). Although each person's situation is different, cancers with the same stage tend to have similar outlooks and are often treated the same way. The cancer stage is also a way for doctors to describe the extent of the cancer when they talk with each other about a person's cancer.",
  "lp_id": "LP6",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "Even though not all new lumps or masses are malignant, they should be checked by a doctor to determine its risks. "
    },
    {
      "id": "SP2",
      "text": "Before taking herbal products, that may help some deal with cancer side effects, patients should consult with their doctor."
    },
    {
      "id": "SP3",
      "text": "The advancement of cancer is categorized and described in stages, which also works as a guide for treatment and prediction."
    },
    {
      "id": "SP4",
      "text": "Doctors usually use stages for cancer so they can predict the evolution and can choose the treatment. Actually they can describe the extension of the tumor through staging."
    },
    {
      "id": "SP5",
      "text": "When doctors are dealing with cancer patients, Oncologists use staging information as a reference to determine the severity of the patient, and their treatment as well.  Even though each patient has their own characteristics, there are similar symptoms during each disease's stage."
    },
    {
      "id": "SP6",
      "text": "Each person with cancer is unique, but cancer stages allow doctors to group similar patients together when they discuss treatment options and outcomes."
    }
  ]
},
{
  "answer": "The most common symptom of breast cancer is a new lump or mass. A painless, hard mass that has irregular edges is more likely to be cancer, but breast cancers can be tender, soft, or round. They can even be painful. For this reason, it's important to have any new breast mass, lump, or breast change checked by an experienced healthcare professional.",
  "lp_id": "LP7",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "The cancer stage represents the advancement state of the cancer. "
    },
```

```
    {
      "id": "SP2",
      "text": "Most breast tumors are benign, rather than malignant (cancer), but needs to be checked by a healthcare professional to be certain."
    },
    {
      "id": "SP3",
      "text": "With the detection of any mass in the breast, it is important to get checked by a doctor."
    },
    {
      "id": "SP4",
      "text": "The first change of breast cancer could be a mass. Pain is not a common or early symptom. Any change you see on your breast, soft or hard, small or big, you should get medical assistance."
    },
    {
      "id": "SP5",
      "text": "A newly devised lump or mass needs to be constantly examined by you and an experienced healthcare professional. It is the most common sign of breast cancer."
    },
    {
      "id": "SP6",
      "text": "If you have a mass or lump in your breast, or a change in the texture, you should talk to your doctor, because these lumps are the most common symptom of breast cancer."
    }
  ]
 },
 {
  "answer": "Lumpectomy is a type of surgery to remove the cancer tumor while saving the nipple and as much of the breast as possible. Depending on how much the cancer has spread, removal of some lymph nodes might also take place during a lumpectomy. A lumpectomy is also called breast conserving surgery.",
  "lp_id": "LP8",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "Any new lump or mass should be checked by a doctor. "
    },
    {
      "id": "SP2",
      "text": "Cancer has different stages and depending on what stage you are on will determine what the plan is for you."
    },
    {
      "id": "SP3",
      "text": "The surgery that cuts off cancer mass and at the same time preserves breast is called lumpectomy "
    },
    {
      "id": "SP4",
      "text": "There is a procedure called lumpectomy where the tumor is extracted from the patient's breast. This procedure may conservate the rest of the breast anatomy allowing the patient to get a normal look after the surgery."
```

```
    },
    {
      "id": "SP5",
      "text": "Lumpectomy is known as breast conservative surgery because it removes the cancer tumor and preserves the nipple and most of the breast tissues. If cancer is spreading it may include some lymph nodes surrounding the area."
    },
    {
      "id": "SP6",
      "text": "A type of surgery that removes a cancerous tumor but saves a lot of the breast tissue and nipple is called a lumpectomy, or breast conserving surgery."
    }
  ]
},
{
  "answer": "A mastectomy is a type of surgery to remove the entire breast. There are several types of mastectomies. Some mastectomies remove all of the breast tissue, including the nipple and areola and the other types of mastectomies also remove some of the nearby lymph nodes",
  "lp_id": "LP9",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "A lumpectomy is surgery to remove the cancer tumor from the breast. "
    },
    {
      "id": "SP2",
      "text": "It is difficult to determine by a new lumps shape or hardness whether it is cancer or not."
    },
    {
      "id": "SP3",
      "text": "There are different types of mastectomy, depending on how much of the breast they remove, up to the complete removal."
    },
    {
      "id": "SP4",
      "text": "Mastectomy is a way to get rid of breast cancer and it would be as radical depending on how far the cancer has spread. The surgeons may take out tissue as long is compromised with the disease."
    },
    {
      "id": "SP5",
      "text": "The mastectomy is a surgery that removes all the breast tissue. There are different types of mastectomy depending on the area removed. The first one focuses on all breast tissues  including the nipple and areola, while others also add nearby lymph nodes to the list."
    },
    {
      "id": "SP6",
      "text": "Mastectomies are surgeries that remove some or all of the breast, and may also remove adjacent lymph nodes."
    }
  ]
},
{
```

```
  "answer": "Systemic therapy refers to medicine that treats the entire body. Systemic therapy for breast cancer is usually given directly into your body through a thin needle or as a pill. It aims to get rid of cancer cells that may have spread from the breast to other parts of the body. Therefore, chemotherapy and targeted therapy reduces the chances of breast cancer coming back. There are three types of systemic therapy: chemotherapy, targeted therapy, and hormonal therapy.",
  "lp_id": "LP10",
  "possible_answers": [
    {
      "id": "SP1",
      "text": "A mastectomy is a very aggressive type of surgery where the entire breast is removed. "
    },
    {
      "id": "SP2",
      "text": "Limiting how much of the nipple is lost and maximizing how much cancer is removed is called Lumpectomy."
    },
    {
      "id": "SP3",
      "text": "The treatment that targets the complete body is called systemic and can be given either directly to the bloodstream or by a pill."
    },
    {
      "id": "SP4",
      "text": "Most types of cancer are treated with systemic therapy, which refers to killing bad cells and avoid the spread of cancer in the rest of the body, minimizing the possibility of relapse. They are called chemotherapy, targeted therapy and hormonal therapy."
    },
    {
      "id": "SP5",
      "text": "Systemic therapy treats your whole body. It enters the body via a needle or as a pill that kills cancer cells that may have spread from the breast to other organs. This reduces the chances of a cancer's come back. There are three types: chemotherapy, targeted therapy and hormonal therapy."
    },
    {
      "id": "SP6",
      "text": "Systemic therapy uses oral medicines, injections, or radiation to try to destroy cancer in the body."
    }
  ]
 }
]
```

APPENDIX B

Cosine similarity values per model, per task (LP1 - LP5)

model	LP1						LP2						LP3						LP4						LP5					
	SP 5	SP 4	SP 3	SP 6	SP 1	SP 2	SP 5	SP 6	SP 4	SP 1	SP 2	SP 3	SP 5	SP 4	SP 3	SP 6	SP 2	SP 1	SP 6	SP 5	SP 4	SP 3	SP 2	SP 1	SP 5	SP 4	SP 2	SP 6	SP 3	SP 1
lsa:- dim:- sco:count sw:yes	0.804	0.666	0.662	0.620	0.517	0.378	0.754	0.749	0.579	0.502	0.449	0.299	0.911	0.707	0.539	0.506	0.494	0.369	0.821	0.818	0.739	0.661	0.611	0.427	0.900	0.897	0.555	0.534	0.517	0.279
lsa:- dim:- sco:count sw:-	0.998	0.996	0.990	0.992	0.980	0.991	0.995	0.993	0.991	0.982	0.995	0.986	0.996	0.993	0.988	0.991	0.982	0.986	0.985	0.996	0.992	0.996	0.994	0.995	0.995	0.997	0.990	0.997	0.996	0.945
lsa:- dim:- sco:tfidf sw:yes	0.656	0.450	0.510	0.480	0.419	0.306	0.678	0.664	0.524	0.393	0.401	0.366	0.770	0.639	0.471	0.455	0.440	0.368	0.693	0.682	0.657	0.572	0.425	0.322	0.675	0.663	0.448	0.482	0.371	0.271
lsa:- dim:- sco:tfidf sw:-	0.993	0.988	0.975	0.975	0.956	0.975	0.986	0.969	0.972	0.953	0.982	0.962	0.984	0.977	0.972	0.972	0.964	0.950	0.955	0.986	0.975	0.984	0.979	0.979	0.987	0.988	0.972	0.990	0.988	0.837
lsa:- dim:- sco:zeroone sw:yes	0.750	0.658	0.695	0.686	0.428	0.253	0.854	0.837	0.722	0.553	0.535	0.392	0.884	0.837	0.608	0.557	0.584	0.438	0.768	0.827	0.788	0.609	0.602	0.537	0.809	0.810	0.666	0.697	0.546	0.325
lsa:- dim:- sco:zeroone sw:-	0.992	0.989	0.977	0.977	0.954	0.964	0.991	0.977	0.975	0.968	0.973	0.968	0.986	0.981	0.969	0.978	0.971	0.959	0.967	0.988	0.983	0.979	0.978	0.975	0.993	0.990	0.982	0.993	0.990	0.938
lsa:yes dim:100 sco:count sw:yes	0.926	0.868	0.786	0.826	0.615	0.407	0.711	0.747	0.491	0.635	0.552	0.244	0.947	0.694	0.739	0.714	0.558	0.346	0.934	0.921	0.856	0.818	0.710	0.458	0.967	0.958	0.645	0.515	0.708	0.282
lsa:yes dim:100 sco:count sw:-	0.998	0.996	0.990	0.991	0.979	0.989	0.994	0.991	0.988	0.979	0.995	0.986	0.996	0.992	0.986	0.990	0.980	0.982	0.984	0.995	0.992	0.995	0.991	0.994	0.993	0.995	0.986	0.997	0.995	0.928
lsa:yes dim:100 sco:tfidf sw:yes	0.829	0.830	0.738	0.725	0.525	0.444	0.729	0.760	0.520	0.635	0.584	0.299	0.934	0.718	0.714	0.819	0.658	0.433	0.932	0.897	0.844	0.818	0.707	0.505	0.949	0.927	0.677	0.567	0.666	0.344
lsa:yes dim:100 sco:tfidf sw:-	0.996	0.991	0.981	0.984	0.962	0.979	0.988	0.979	0.976	0.963	0.987	0.970	0.993	0.983	0.977	0.984	0.969	0.965	0.973	0.990	0.981	0.987	0.984	0.984	0.988	0.991	0.977	0.993	0.991	0.847
lsa:yes dim:100 sco:zeroone sw:yes	0.923	0.924	0.918	0.881	0.824	0.700	0.899	0.897	0.768	0.846	0.795	0.720	0.964	0.894	0.819	0.904	0.868	0.803	0.916	0.926	0.879	0.877	0.759	0.895	0.955	0.918	0.785	0.847	0.892	0.701
lsa:yes dim:100 sco:zeroone sw:-	0.997	0.994	0.992	0.986	0.976	0.974	0.996	0.988	0.982	0.980	0.980	0.979	0.992	0.991	0.979	0.992	0.983	0.984	0.978	0.992	0.990	0.987	0.988	0.984	0.997	0.996	0.988	0.997	0.994	0.985
lsa:yes dim:300 sco:count sw:yes	0.854	0.642	0.698	0.707	0.490	0.362	0.681	0.714	0.483	0.608	0.562	0.288	0.928	0.678	0.785	0.685	0.492	0.260	0.884	0.874	0.808	0.749	0.649	0.430	0.950	0.941	0.564	0.504	0.651	0.297
lsa:yes dim:300 sco:count sw:-	0.998	0.995	0.990	0.991	0.979	0.989	0.994	0.991	0.988	0.978	0.995	0.985	0.996	0.992	0.986	0.990	0.979	0.982	0.981	0.995	0.992	0.995	0.991	0.994	0.993	0.995	0.986	0.997	0.995	0.930
lsa:yes dim:300 sco:tfidf sw:yes	0.802	0.800	0.627	0.701	0.473	0.308	0.673	0.631	0.452	0.580	0.548	0.332	0.898	0.671	0.614	0.676	0.453	0.240	0.883	0.824	0.751	0.780	0.560	0.449	0.904	0.865	0.583	0.552	0.634	0.386
lsa:yes dim:300 sco:tfidf sw:-	0.995	0.990	0.980	0.981	0.962	0.978	0.987	0.974	0.974	0.959	0.985	0.968	0.988	0.979	0.973	0.981	0.964	0.958	0.964	0.988	0.978	0.986	0.981	0.982	0.988	0.990	0.974	0.993	0.990	0.837
lsa:yes dim:300 sco:zeroone sw:yes	0.832	0.725	0.773	0.742	0.692	0.562	0.862	0.829	0.707	0.713	0.647	0.589	0.925	0.865	0.736	0.842	0.719	0.717	0.824	0.849	0.811	0.789	0.682	0.779	0.875	0.876	0.689	0.745	0.582	0.620
lsa:yes dim:300 sco:zeroone sw:-	0.995	0.989	0.986	0.981	0.966	0.971	0.993	0.982	0.979	0.973	0.977	0.973	0.987	0.985	0.972	0.983	0.976	0.974	0.969	0.988	0.985	0.982	0.981	0.978	0.995	0.992	0.984	0.995	0.991	0.947

Cosine similarity values per model, per task (LP6 - LP10)

model	LP6						LP7						LP8						LP9						LP10					
	SP 5	SP 4	SP 2	SP 1	SP 6	SP 3	SP 4	SP 3	SP 6	SP 1	SP 5	SP 2	SP 5	SP 3	SP 2	SP 6	SP 4	SP 1	SP 5	SP 4	SP 3	SP 2	SP 6	SP 1	SP 6	SP 3	SP 5	SP 4	SP 2	SP 1
lsa:- dim:- sco:count sw:yes	0.888	0.615	0.602	0.579	0.526	0.384	0.656	0.652	0.630	0.548	0.538	0.498	0.891	0.657	0.612	0.591	0.409	0.246	0.694	0.520	0.477	0.460	0.409	0.388	0.752	0.746	0.638	0.621	0.476	0.414
lsa:- dim:- sco:count sw:-	0.996	0.986	0.966	0.986	0.994	0.994	0.996	0.997	0.996	0.981	0.997	0.992	0.990	0.996	0.993	0.996	0.995	0.975	0.996	0.996	0.996	0.984	0.997	0.994	0.993	0.998	0.997	0.995	0.993	0.996
lsa:- dim:- sco:tfidf sw:yes	0.720	0.499	0.463	0.356	0.462	0.335	0.612	0.574	0.668	0.462	0.551	0.488	0.827	0.634	0.526	0.533	0.410	0.130	0.548	0.385	0.356	0.301	0.413	0.394	0.640	0.641	0.569	0.533	0.305	0.286
lsa:- dim:- sco:tfidf sw:-	0.987	0.969	0.905	0.962	0.977	0.985	0.984	0.989	0.983	0.954	0.990	0.973	0.968	0.983	0.978	0.986	0.982	0.921	0.985	0.980	0.985	0.955	0.990	0.977	0.980	0.991	0.993	0.984	0.979	0.986
lsa:- dim:- sco:zeroone sw:yes	0.861	0.716	0.709	0.645	0.706	0.497	0.779	0.726	0.753	0.640	0.669	0.597	0.601	0.698	0.551	0.705	0.668	0.488	0.665	0.468	0.414	0.452	0.288	0.302	0.746	0.782	0.729	0.698	0.453	0.401
lsa:- dim:- sco:zeroone sw:-	0.988	0.979	0.947	0.977	0.981	0.979	0.983	0.979	0.981	0.937	0.982	0.981	0.971	0.978	0.979	0.979	0.977	0.953	0.982	0.973	0.973	0.968	0.986	0.962	0.973	0.985	0.994	0.980	0.964	0.978
lsa:yes dim:100 sco:count sw:yes	0.945	0.690	0.722	0.569	0.691	0.521	0.659	0.794	0.684	0.509	0.622	0.600	0.963	0.675	0.680	0.722	0.424	0.301	0.920	0.806	0.850	0.688	0.410	0.379	0.889	0.910	0.751	0.782	0.775	0.717
lsa:yes dim:100 sco:count sw:-	0.996	0.986	0.959	0.984	0.992	0.993	0.996	0.997	0.997	0.983	0.997	0.990	0.988	0.995	0.995	0.995	0.994	0.969	0.995	0.995	0.995	0.981	0.997	0.993	0.992	0.998	0.996	0.995	0.993	0.996
lsa:yes dim:100 sco:tfidf sw:yes	0.900	0.768	0.775	0.535	0.739	0.569	0.690	0.782	0.710	0.543	0.710	0.652	0.958	0.559	0.699	0.721	0.429	0.219	0.916	0.828	0.816	0.649	0.455	0.432	0.892	0.902	0.738	0.761	0.716	0.675
lsa:yes dim:100 sco:tfidf sw:-	0.992	0.975	0.944	0.969	0.987	0.988	0.991	0.994	0.991	0.963	0.994	0.982	0.971	0.987	0.985	0.989	0.988	0.943	0.988	0.986	0.989	0.963	0.993	0.981	0.983	0.995	0.994	0.989	0.985	0.990
lsa:yes dim:100 sco:zeroone sw:yes	0.939	0.886	0.896	0.752	0.928	0.880	0.867	0.793	0.858	0.702	0.878	0.868	0.882	0.795	0.912	0.906	0.854	0.674	0.940	0.749	0.868	0.760	0.843	0.678	0.961	0.953	0.854	0.935	0.851	0.889
lsa:yes dim:100 sco:zeroone sw:-	0.996	0.988	0.969	0.986	0.993	0.985	0.991	0.984	0.989	0.958	0.987	0.994	0.985	0.988	0.991	0.988	0.989	0.962	0.992	0.985	0.989	0.985	0.994	0.975	0.988	0.991	0.998	0.994	0.991	0.989
lsa:yes dim:300 sco:count sw:yes	0.907	0.620	0.647	0.548	0.644	0.479	0.580	0.748	0.642	0.515	0.622	0.606	0.900	0.690	0.662	0.704	0.458	0.344	0.891	0.707	0.833	0.650	0.441	0.393	0.886	0.853	0.721	0.726	0.631	0.645
lsa:yes dim:300 sco:count sw:-	0.996	0.986	0.958	0.984	0.992	0.993	0.996	0.997	0.996	0.982	0.997	0.990	0.988	0.995	0.994	0.995	0.994	0.968	0.995	0.995	0.995	0.981	0.997	0.993	0.992	0.998	0.996	0.995	0.993	0.995
lsa:yes dim:300 sco:tfidf sw:yes	0.803	0.655	0.664	0.435	0.689	0.542	0.607	0.722	0.648	0.477	0.697	0.659	0.911	0.555	0.548	0.677	0.488	0.265	0.822	0.667	0.734	0.604	0.478	0.436	0.882	0.875	0.698	0.783	0.618	0.628
lsa:yes dim:300 sco:tfidf sw:-	0.990	0.973	0.914	0.963	0.984	0.987	0.987	0.990	0.986	0.959	0.992	0.979	0.970	0.986	0.982	0.987	0.985	0.923	0.987	0.982	0.986	0.956	0.992	0.980	0.982	0.993	0.994	0.987	0.984	0.989
lsa:yes dim:300 sco:zeroone sw:yes	0.900	0.831	0.785	0.710	0.845	0.785	0.827	0.762	0.743	0.671	0.722	0.605	0.649	0.735	0.635	0.745	0.766	0.562	0.798	0.489	0.484	0.489	0.553	0.457	0.906	0.845	0.775	0.851	0.517	0.672
lsa:yes dim:300 sco:zeroone sw:-	0.991	0.985	0.956	0.979	0.988	0.982	0.985	0.979	0.983	0.947	0.983	0.984	0.974	0.982	0.982	0.982	0.979	0.954	0.986	0.976	0.979	0.973	0.990	0.965	0.984	0.988	0.997	0.990	0.981	0.982

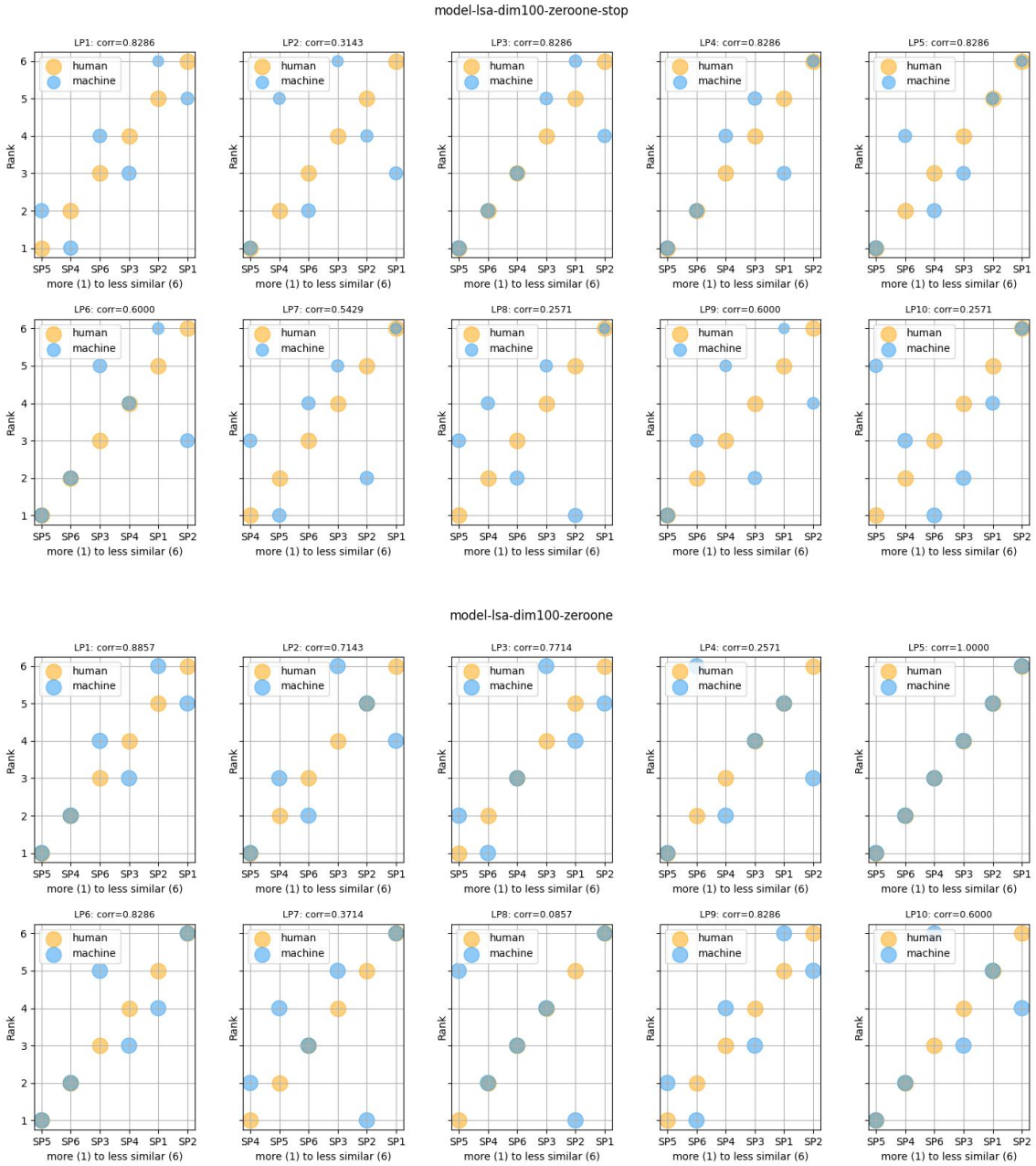
Graphs from all 18 models, showing human and machine ranking results.


model-docbyterm-tfidf-stop
LP1: corr=0.7143
LP2: corr=0.7714
LP3: corr=0.7714
LP4: corr=0.8857
LP5: corr=0.8857
LP6: corr=0.2571
LP7: corr=0.7143
LP8: corr=0.6000
LP9: corr=0.8286
LP10: corr=0.4286
model-docbyterm-tfidf
LP1: corr=0.7714
LP2: corr=0.6571
LP3: corr=0.7714
LP4: corr=0.0286
LP5: corr=0.6571
LP6: corr=0.9429
LP7: corr=0.7143
LP8: corr=0.2571
LP9: corr=0.7143
LP10: corr=0.5429
human
machine
Rank
more (1) to less similar (6)


model-docbyterm-zeroone-stop
LP1: corr=0.7143
LP2: corr=0.7143
LP3: corr=0.5429
LP4: corr=0.8857
LP5: corr=0.7714
LP6: corr=0.2571
LP7: corr=0.7714
LP8: corr=0.4857
LP9: corr=0.2571
LP10: corr=0.4286
model-docbyterm-zeroone
LP1: corr=1.0000
LP2: corr=0.7714
LP3: corr=0.7714
LP4: corr=0.3714
LP5: corr=0.9429
LP6: corr=1.0000
LP7: corr=0.9429
LP8: corr=-0.0857
LP9: corr=0.8286
LP10: corr=0.7143
human
machine
Rank
more (1) to less similar (6)


model-lsa-dim100-count-stop
LP1: corr=0.9429
LP2: corr=0.2000
LP3: corr=0.7714
LP4: corr=0.8857
LP5: corr=0.6571
LP6: corr=0.2571
LP7: corr=0.4857
LP8: corr=0.6000
LP9: corr=0.4857
LP10: corr=0.0857
model-lsa-dim100-count
LP1: corr=1.0000
LP2: corr=0.3714
LP3: corr=0.9429
LP4: corr=0.0857
LP5: corr=0.6000
LP6: corr=0.9429
LP7: corr=0.6571
LP8: corr=0.0286
LP9: corr=0.6000
LP10: corr=0.2000
human
machine
Rank
more (1) to less similar (6)


model-lsa-dim100-tfidf-stop
LP1: corr=0.8286
LP2: corr=0.2000
LP3: corr=0.9429
LP4: corr=0.8857
LP5: corr=0.6000
LP6: corr=0.2571
LP7: corr=0.4857
LP8: corr=0.6000
LP9: corr=0.5429
LP10: corr=0.3714
model-lsa-dim100-tfidf
LP1: corr=1.0000
LP2: corr=0.6000
LP3: corr=0.9429
LP4: corr=0.0286
LP5: corr=0.6571
LP6: corr=0.9429
LP7: corr=0.6000
LP8: corr=0.3714
LP9: corr=0.7143
LP10: corr=0.2000
human
machine
Rank
more (1) to less similar (6)


model-lsa-dim100-zeroone-stop
LP1: corr=0.8286
LP2: corr=0.3143
LP3: corr=0.8286
LP4: corr=0.8286
LP5: corr=0.8286
LP6: corr=0.6000
LP7: corr=0.5429
LP8: corr=0.2571
LP9: corr=0.6000
LP10: corr=0.2571
model-lsa-dim100-zeroone
LP1: corr=0.8857
LP2: corr=0.7143
LP3: corr=0.7714
LP4: corr=0.2571
LP5: corr=1.0000
LP6: corr=0.8286
LP7: corr=0.3714
LP8: corr=0.0857
LP9: corr=0.8286
LP10: corr=0.6000
human
machine
Rank
more (1) to less similar (6)


model-lsa-dim300-count-stop
LP1: corr=0.7714
LP2: corr=0.2000
LP3: corr=0.8857
LP4: corr=0.8857
LP5: corr=0.6571
LP6: corr=0.2571
LP7: corr=0.2571
LP8: corr=0.6571
LP9: corr=0.4857
LP10: corr=0.4857
model-lsa-dim300-count
LP1: corr=1.0000
LP2: corr=0.3714
LP3: corr=0.9429
LP4: corr=0.0857
LP5: corr=0.6000
LP6: corr=0.9429
LP7: corr=0.6000
LP8: corr=0.0286
LP9: corr=0.6000
LP10: corr=0.2000
human
machine
Rank
more (1) to less similar (6)


model-lsa-dim300-tfidf-stop
LP1: corr=0.9429
LP2: corr=0.3143
LP3: corr=0.9429
LP4: corr=0.8286
LP5: corr=0.6571
LP6: corr=0.6000
LP7: corr=0.2571
LP8: corr=0.6571
LP9: corr=0.4857
LP10: corr=0.4857
model-lsa-dim300-tfidf
LP1: corr=1.0000
LP2: corr=0.6571
LP3: corr=0.9429
LP4: corr=0.0857
LP5: corr=0.6000
LP6: corr=0.9429
LP7: corr=0.7143
LP8: corr=0.2571
LP9: corr=0.8857
LP10: corr=0.3714
human
machine
Rank
more (1) to less similar (6)

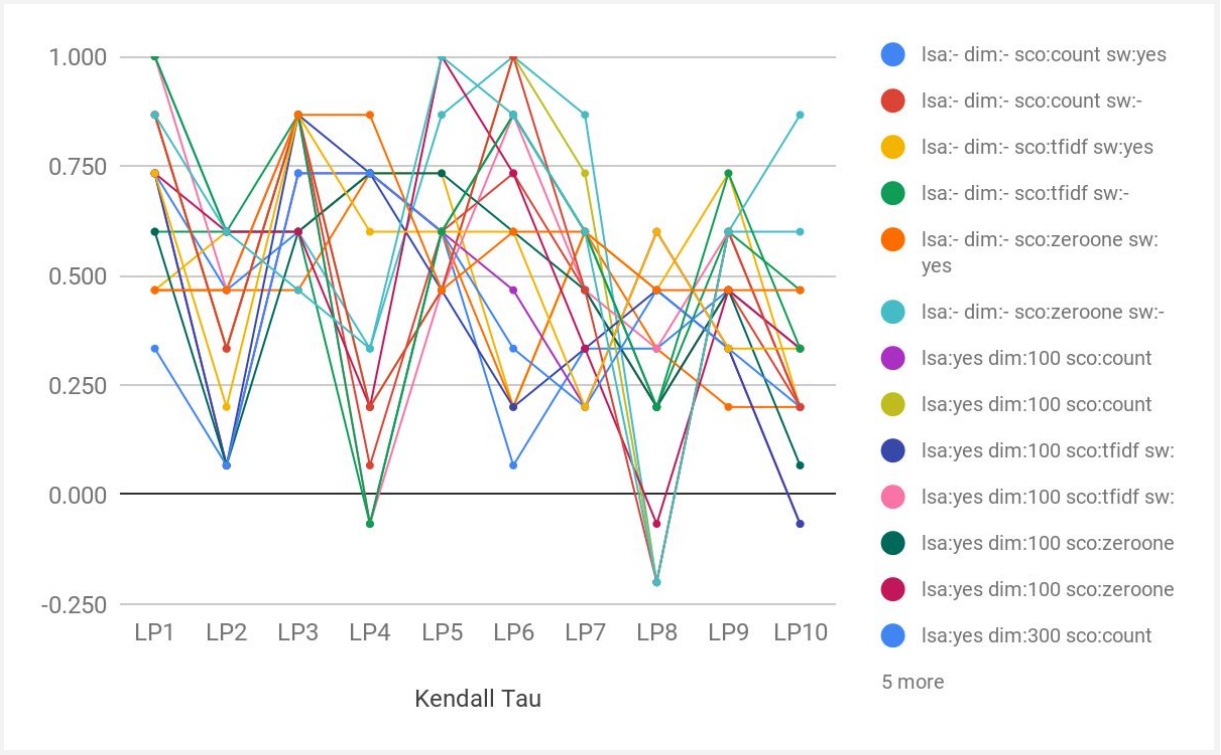

model-lsa-dim300-zeroone-stop
LP1: corr=0.7143
LP2: corr=0.5429
LP3: corr=0.8857
LP4: corr=1.0000
LP5: corr=0.6571
LP6: corr=0.8857
LP7: corr=0.7143
LP8: corr=0.6571
LP9: corr=0.8286
LP10: corr=0.6000
model-lsa-dim300-zeroone
LP1: corr=0.9429
LP2: corr=0.8857
LP3: corr=0.7143
LP4: corr=0.3714
LP5: corr=1.0000
LP6: corr=0.9429
LP7: corr=0.6571
LP8: corr=-0.0857
LP9: corr=0.8286
LP10: corr=0.9429
human
machine
Rank
more (1) to less similar (6)

APPENDIX C

Kendall Tau Correlation Table

Kendall Tau	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8	LP9	LP10
lsa:- dim:- sco:count sw:yes	0.733	0.467	0.600	0.733	0.600	0.067	0.333	0.333	0.467	0.333
lsa:- dim:- sco:count sw:-	0.867	0.333	0.867	0.067	0.600	0.733	0.467	0.200	0.467	0.200
lsa:- dim:- sco:tfidf sw:yes	0.467	0.600	0.600	0.733	0.733	0.200	0.600	0.467	0.733	0.200
lsa:- dim:- sco:tfidf sw:-	0.600	0.600	0.600	-0.067	0.600	0.867	0.600	0.200	0.600	0.467
lsa:- dim:- sco:zeroone sw:yes	0.467	0.467	0.467	0.733	0.600	0.200	0.600	0.333	0.200	0.200
lsa:- dim:- sco:zeroone sw:-	1.000	0.600	0.600	0.333	0.867	1.000	0.867	-0.200	0.600	0.600
lsa:yes dim:100 sco:count sw:yes	0.733	0.067	0.733	0.733	0.600	0.467	0.200	0.600	0.333	-0.067
lsa:yes dim:100 sco:count sw:-	0.867	0.333	0.867	0.200	0.467	1.000	0.733	-0.200	0.600	0.200
lsa:yes dim:100 sco:tfidf sw:yes	0.733	0.067	0.867	0.733	0.467	0.200	0.333	0.467	0.333	-0.067
lsa:yes dim:100 sco:tfidf sw:-	1.000	0.467	0.867	-0.067	0.467	0.867	0.467	0.333	0.600	0.200
lsa:yes dim:100 sco:zeroone sw:yes	0.600	0.067	0.600	0.733	0.733	0.600	0.467	0.200	0.467	0.067
lsa:yes dim:100 sco:zeroone sw:-	0.733	0.600	0.600	0.200	1.000	0.733	0.333	-0.067	0.467	0.333
lsa:yes dim:300 sco:count sw:yes	0.333	0.067	0.733	0.733	0.600	0.333	0.200	0.467	0.333	0.200
lsa:yes dim:300 sco:count sw:-	0.867	0.333	0.867	0.200	0.467	1.000	0.467	-0.200	0.600	0.200
lsa:yes dim:300 sco:tfidf sw:yes	0.733	0.200	0.867	0.600	0.600	0.600	0.200	0.600	0.333	0.333
lsa:yes dim:300 sco:tfidf sw:-	1.000	0.600	0.867	-0.067	0.600	0.867	0.600	0.200	0.733	0.333
lsa:yes dim:300 sco:zeroone sw:yes	0.467	0.467	0.867	0.867	0.467	0.600	0.600	0.467	0.467	0.467
lsa:yes dim:300 sco:zeroone sw:-	0.867	0.600	0.467	0.333	1.000	0.867	0.600	-0.200	0.600	0.867

Kendall Tau Correlation



Order Preservation Measure

Order Preservation Measure	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8	LP9	LP10
lsa:- dim:- sco:count sw:yes	0.867	0.733	0.800	0.867	0.800	0.533	0.667	0.667	0.733	0.667
lsa:- dim:- sco:count sw:-	0.933	0.667	0.933	0.533	0.800	0.867	0.733	0.600	0.733	0.600
lsa:- dim:- sco:tfidf sw:yes	0.733	0.800	0.800	0.867	0.867	0.600	0.800	0.733	0.867	0.600
lsa:- dim:- sco:tfidf sw:-	0.800	0.800	0.800	0.467	0.800	0.933	0.800	0.600	0.800	0.733
lsa:- dim:- sco:zeroone sw:yes	0.733	0.733	0.733	0.867	0.800	0.600	0.800	0.667	0.600	0.600
lsa:- dim:- sco:zeroone sw:-	1.000	0.800	0.800	0.667	0.933	1.000	0.933	0.400	0.800	0.800
lsa:yes dim:100 sco:count sw:yes	0.867	0.533	0.867	0.867	0.800	0.733	0.600	0.800	0.667	0.467
lsa:yes dim:100 sco:count sw:-	0.933	0.667	0.933	0.600	0.733	1.000	0.867	0.400	0.800	0.600
lsa:yes dim:100 sco:tfidf sw:yes	0.867	0.533	0.933	0.867	0.733	0.600	0.667	0.733	0.667	0.467
lsa:yes dim:100 sco:tfidf sw:-	1.000	0.733	0.933	0.467	0.733	0.933	0.733	0.667	0.800	0.600
lsa:yes dim:100 sco:zeroone sw:yes	0.800	0.533	0.800	0.867	0.867	0.800	0.733	0.600	0.733	0.533
lsa:yes dim:100 sco:zeroone sw:-	0.867	0.800	0.800	0.600	1.000	0.867	0.667	0.467	0.733	0.667
lsa:yes dim:300 sco:count sw:yes	0.667	0.533	0.867	0.867	0.800	0.667	0.600	0.733	0.667	0.600
lsa:yes dim:300 sco:count sw:-	0.933	0.667	0.933	0.600	0.733	1.000	0.733	0.400	0.800	0.600
lsa:yes dim:300 sco:tfidf sw:yes	0.867	0.600	0.933	0.800	0.800	0.800	0.600	0.800	0.667	0.667
lsa:yes dim:300 sco:tfidf sw:-	1.000	0.800	0.933	0.467	0.800	0.933	0.800	0.600	0.867	0.667

lsa:yes dim:300 sco:zeroone sw:yes	0.733	0.733	0.933	0.933	0.733	0.800	0.800	0.733	0.733	0.733
lsa:yes dim:300 sco:zeroone sw:-	0.933	0.800	0.733	0.667	1.000	0.933	0.800	0.400	0.800	0.933